BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e38263
doi: 10.3897/biss.3.38263

OPEN ACCESS

Conference Abstract

Citing Evolving Data: An implementation on the NHM Data Portal

Josh Humphries ‡

‡ Natural History Museum, London, United Kingdom

Corresponding author: Josh Humphries (j.humphries@nhm.ac.uk)

Received: 15 Jul 2019 | Published: 17 Jul 2019

Citation: Humphries J (2019) Citing Evolving Data: An implementation on the NHM Data Portal. Biodiversity Information Science and Standards 3: e38263. https://doi.org/10.3897/biss.3.38263

Abstract

Since 2015, the Natural History Museum London has made its research and collections data available through its Data Portal (https://data.nhm.ac.uk). This website provides free and open access to important research datasets as well as digitised objects from the Museum's specimen collection. The Data Portal currently has over 4.2 million records from the specimen collection and a further 5.5 million records from other research datasets. Since 2015, more than 250 scientific publications have cited data from the Data Portal, either directly or through aggregators such as the Global Biodiversity Information Facility (GBIF), although there are many more citations than it is currently possible to track.

Users can download data from the Portal and are encouraged to cite the source, however, there is currently no way for users to cite subsets of the data returned through a query, nor a way to persistently identify the data subset they are citing. This is a common issue with scientific data put online, particularly when the cited data changes frequently, such as is the case with the Museum's specimen collection, which grows constantly as more of the collection is digitised.

This poster outlines a new approach that has been designed to meet the Research Data Alliance's (RDA) Working Group on Data Citation recommendations on citing evolving data (Rauber et al. 2015). This is achieved by implementing a fully versioned search framework, ensuring that all modifications to records are tracked and the version timestamp of each modification is combined with the data into the search index. When users search and

download data from the Portal, Digital Object Identifiers (DOI) are minted for unique searches at exact versions allowing the dynamic, repeated retrieval of data at any version timestamp, without storing the results. Combining the versioning information into the search index also allows queries against historical data.

By persistently identifying query results in this fashion, researchers can cite data precisely and have confidence that although the data may change after they use it, users of their work will be able to access the data as it looked when they studied it originally. This should also encourage the systematic use of citations, making it easier to track both the usage and impact of research and collections datasets.

Keywords

data portal, citation, DOI, persistent identifiers

Presenting author

Josh Humphries

Presented at

Biodiversity_Next 2019

Funding program

Supported by the Natural History Museum, London.

Author contributions

Alice Butcher, Matt Woodburn

References

- Rauber A, Asmi A, Uytvanck Dv, Proell S (2015) Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). https://doi.org/10.15497/RDA00016. Accessed on: 2019-4-05.